

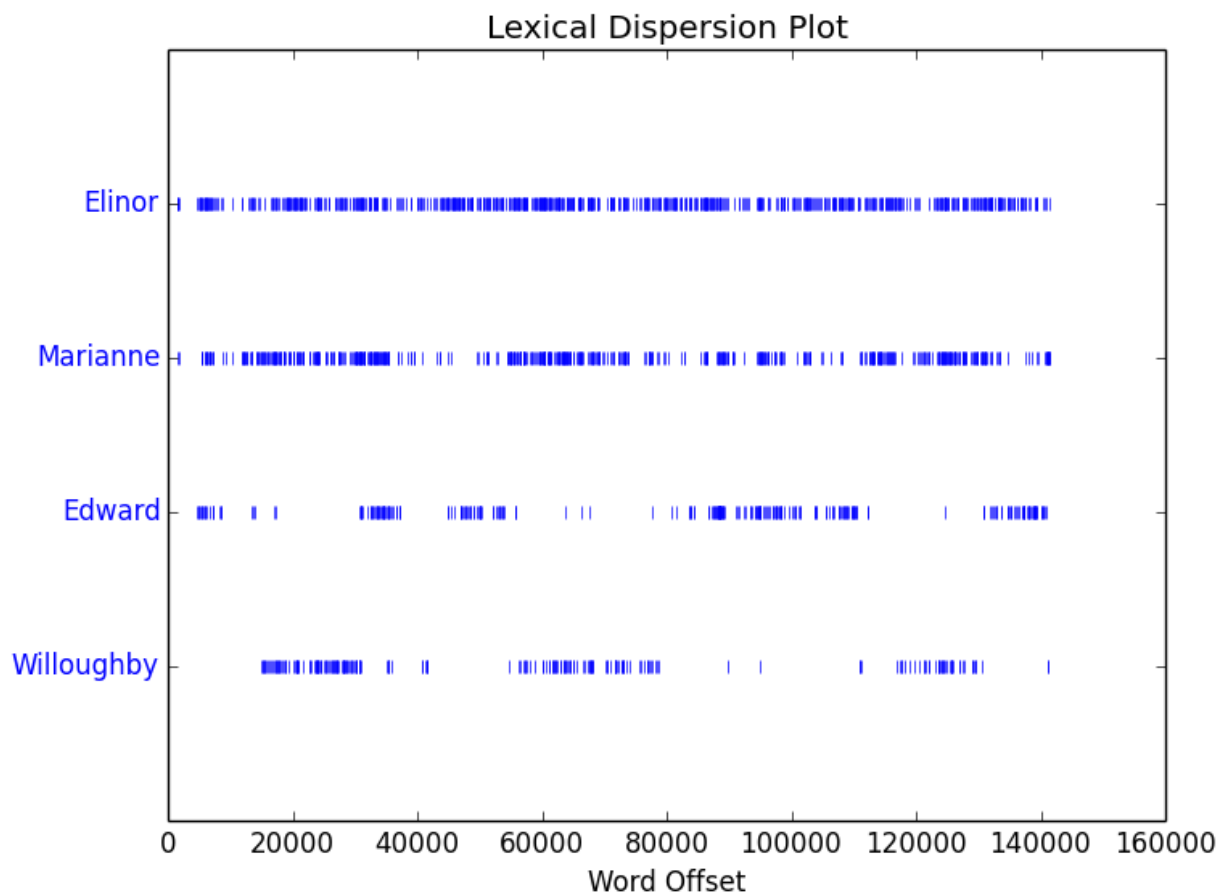
中文信息处理 Chinese Information Processing

第一章 作业

14307130318 刘超颖

1. 制作text2 (《理智与情感》) 中四个主角: Elinor, Marianne, Edward 和Willoughby 的分布图。在这部小说中关于男性和女性所扮演的不同角色, 你能观察到什么?

```
text2.dispersion_plot(["Elinor", "Marianne", "Edward", "Willoughby"])
```

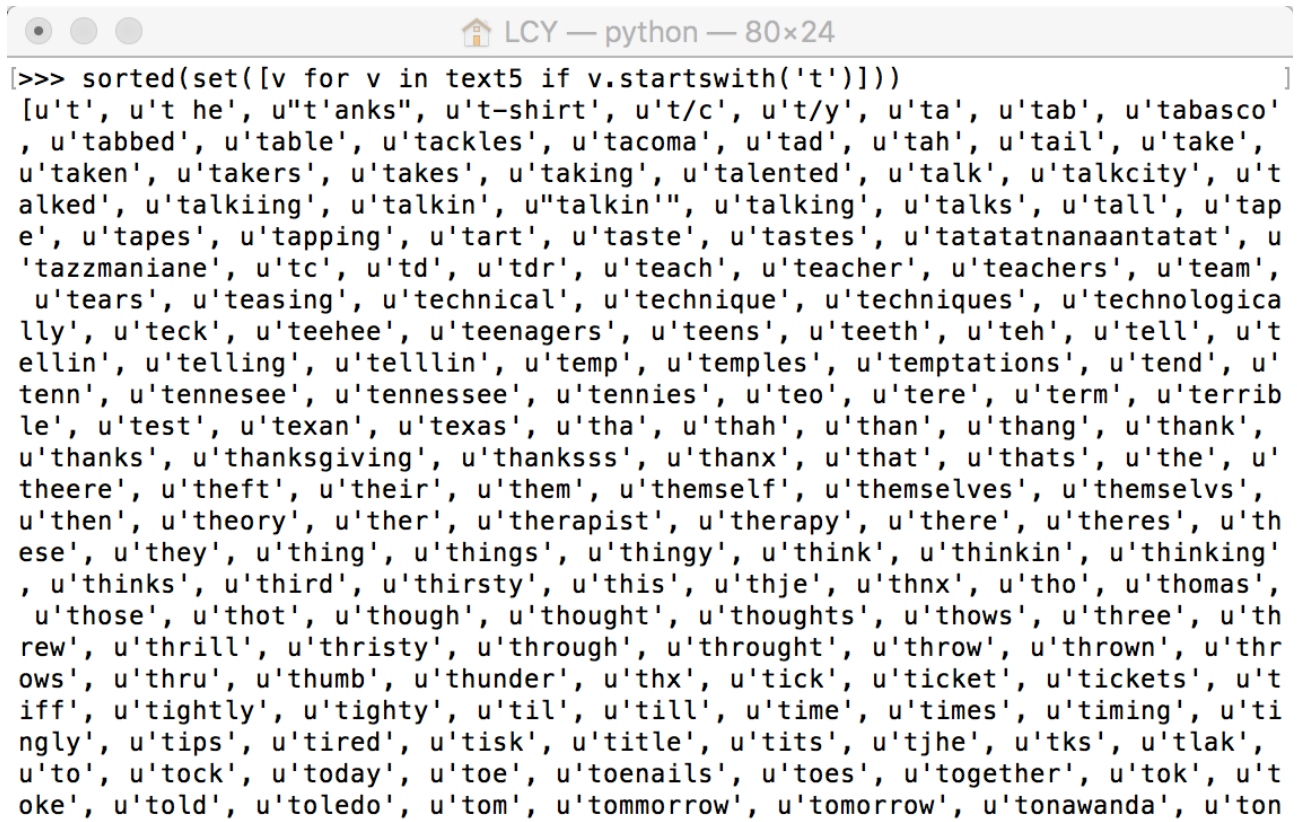


Elinor和Marianne是本书的女主角, Edward和Willoughby是本书的男主角。女主角出现频率更高, 说明《理智与情感》是一本女性向的小说, 从女主角角度叙述故事。Elinor和Edward是一对情侣, Marianne和Willoughby是一对情侣, 因为他们的名字总是同时出现。

2. 在聊天语料库 (text5) 中查找所有以字母t开头的词，按字母顺序显示出来。找出text5中所有5个字母的词。使用频率分布函数 (FreqDist) ，以频率从高到低显示这些词。

(1) 在聊天语料库 (text5) 中查找所有以字母t开头的词，按字母顺序显示出来。

```
sorted(set([v for v in text5 if v.startswith('t'))))
```



```
LCY — python — 80x24

[>>> sorted(set([v for v in text5 if v.startswith('t'))))
[u't', u't he', u't'anks", u't-shirt', u't/c', u't/y', u'ta', u'tab', u'tabasco',
u'tabbed', u'table', u'tackles', u'tacoma', u'tad', u'tah', u'tail', u'take',
u'taken', u'takers', u'takes', u'taking', u'talented', u'talk', u'talkcity', u't
alked', u'talkiing', u'talkin', u'talkin'", u'talking', u'talks', u'tall', u'tap
e', u'tapes', u'tapping', u'tart', u'taste', u'tastes', u'tatatatnanaantatat', u
'tazzmaniane', u'tc', u'td', u'tdr', u'teach', u'teacher', u'teachers', u'team',
u'tears', u'teasing', u'technical', u'technique', u'techniques', u'technologica
lly', u'teck', u'teehee', u'teenagers', u'teens', u'teeth', u'teh', u'tell', u't
ellin', u'telling', u'telllin', u'temp', u'temples', u'temptations', u'tend', u'
tenn', u'tennessee', u'tennessee', u'tennies', u'teo', u'tere', u'term', u'terrib
le', u'test', u'texan', u'texas', u'tha', u'thah', u'than', u'thang', u'thank',
u'thanks', u'thanksgiving', u'thanksss', u'thanx', u'that', u'thats', u'the', u'
theere', u'theft', u'their', u'them', u'themself', u'themselves', u'themselvs',
u'then', u'theory', u'ther', u'therapist', u'therapy', u'there', u'theres', u'th
ese', u'they', u'thing', u'things', u'thingy', u'think', u'thinkin', u'thinking',
u'thinks', u'third', u'thirsty', u'this', u'thje', u'thnx', u'tho', u'thomas',
u'those', u'thot', u'though', u'thought', u'thoughts', u'thows', u'three', u'th
rew', u'thrill', u'thristy', u'through', u'throught', u'throw', u'thrown', u'thr
ows', u'thru', u'thumb', u'thunder', u'thx', u'tick', u'ticket', u'tickets', u't
iff', u'tightly', u'tighty', u'til', u'till', u'time', u'times', u'timing', u'ti
ngly', u'tips', u'tired', u'tisk', u'title', u'tits', u'tjhe', u'tks', u'tlak',
u'to', u'tock', u'today', u'toe', u'toenails', u'toes', u'together', u'tok', u't
oke', u'told', u'toledo', u'tom', u'tomorrow', u'tomorrow', u'tonawanda', u'ton
```

(2) 找出text5中所有5个字母的词。使用频率分布函数 (FreqDist) , 以频率从高到低显示这些词。

```
fdist = FreqDist([w for w in text5 if len(w) ==5])
fdist.most_common()
```

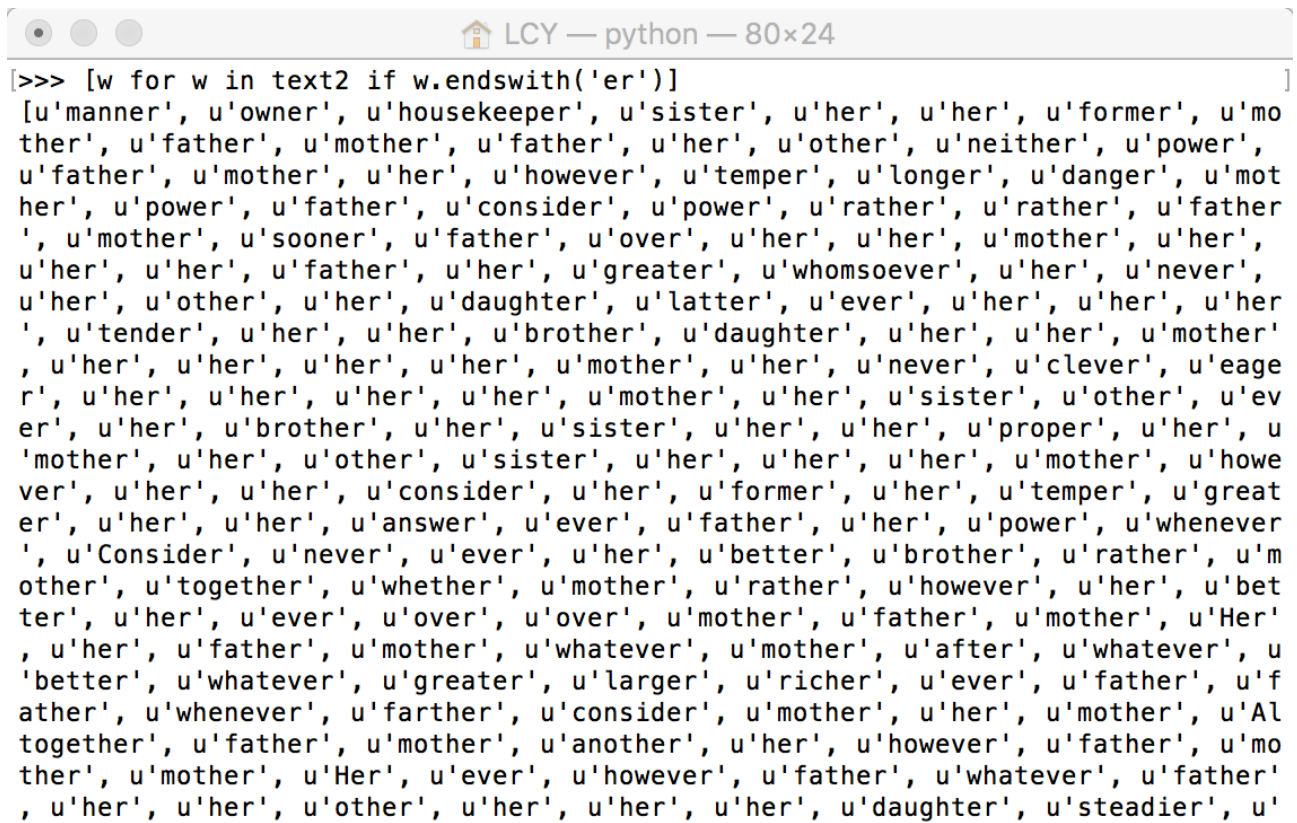
```
LCY — python — 80×24

>>> fdist = FreqDist([w for w in text5 if len(w) == 5])
>>> fdist.most_common()
[(u'there', 120), (u'wanna', 107), (u'.....', 73), (u'hello', 71), (u'about', 70), (u'where', 63), (u'think', 54), (u'right', 54), (u'would', 53), (u'girls', 48), (u'never', 45), (u'thats', 45), (u'whats', 41), (u'night', 41), (u'gonna', 37), (u'still', 33), (u'today', 29), (u'sorry', 28), (u'didnt', 28), (u'going', 27), (u'again', 23), (u'first', 22), (u'looks', 21), (u'doing', 21), (u'guess', 21), (u'wants', 21), (u'great', 20), (u'phone', 20), (u'could', 20), (u'maybe', 20), (u'bored', 20), (u'howdy', 19), (u'their', 19), (u'later', 19), (u'other', 19), (u'sucks', 18), (u'thing', 18), (u'Hello', 18), (u'leave', 17), (u'tryin', 17), (u'thank', 14), (u'hands', 14), (u'least', 14), (u''))), (u'these', 14), (u'makes', 13), (u'music', 13), (u'!!!!', 13), (u'cause', 13), (u'wrong', 13), (u'times', 13), (u'place', 13), (u'games', 13), (u'sleep', 12), (u'gotta', 12), (u'outta', 12), (u'watch', 12), (u'naked', 12), (u'sweet', 12), (u'those', 12), (u'party', 12), (u'being', 12), (u'honey', 11), (u'stuff', 11), (u'names', 10), (u'Music', 10), (u'hahah', 10), (u'cream', 10), (u'alone', 9), (u'hiYas', 9), (u'waves', 9), (u'white', 9), (u'meant', 9), (u'while', 9), (u'funny', 9), (u'aloha', 9), (u'(((((', 9), (u'lasts', 9), (u'happy', 9), (u'stick', 8), (u'heard', 8), (u'tried', 8), (u'after', 8), (u'years', 8), (u'jesus', 8), (u'black', 8), (u'point', 8), (u'lived', 8), (u'world', 8), (u'sighs', 8), (u'trout', 8), (u'kinda', 8), (u'gurls', 8), (u'kicks', 7), (u'video', 7), (u'knows', 7), (u'women', 7), (u'beach', 7), (u'short', 7), (u'green', 7), (u'since', 7), (u'dunno', 7), (u'Paxil', 7), (u'loves', 7), (u'couch', 7), (u'folks', 7), (u'drink', 7), (u'seems', 7), (u'?????', 7), (u'moped', 7), (u'agree', 7), (u'walks', 7), (u'check'
```

3. 写表达式找出text2 中所有符合下列条件的词。结果应该是词链表的形式: ['word 1', 'word2', ...]。a. 以er 结尾; b. 包含字母m; c. 包含字母序列ph; d. 除了首字母外是全部小写字母的词 (即titlecase) 。

(a) 以er 结尾

```
[w for w in text2 if w.endswith('er')]
```



The screenshot shows a terminal window titled "LCY — python — 80x24". The prompt is ">>>". The code entered is "[w for w in text2 if w.endswith('er')]" followed by a closing bracket. The output is a long list of words in single quotes, all ending in 'er'. The words are: 'manner', 'owner', 'housekeeper', 'sister', 'her', 'her', 'former', 'mother', 'father', 'mother', 'father', 'her', 'other', 'neither', 'power', 'father', 'mother', 'her', 'however', 'temper', 'longer', 'danger', 'mother', 'power', 'father', 'consider', 'power', 'rather', 'rather', 'father', 'mother', 'sooner', 'father', 'over', 'her', 'her', 'mother', 'her', 'her', 'her', 'father', 'her', 'greater', 'whomsoever', 'her', 'never', 'her', 'other', 'her', 'daughter', 'latter', 'ever', 'her', 'her', 'her', 'tender', 'her', 'her', 'brother', 'daughter', 'her', 'her', 'mother', 'her', 'her', 'her', 'her', 'mother', 'her', 'never', 'clever', 'eager', 'her', 'her', 'her', 'her', 'mother', 'her', 'sister', 'other', 'ever', 'her', 'brother', 'her', 'sister', 'her', 'her', 'proper', 'her', 'mother', 'her', 'other', 'sister', 'her', 'her', 'her', 'mother', 'however', 'her', 'her', 'consider', 'her', 'former', 'her', 'temper', 'greater', 'her', 'her', 'answer', 'ever', 'father', 'her', 'power', 'whenever', 'Consider', 'never', 'ever', 'her', 'better', 'brother', 'rather', 'mother', 'together', 'whether', 'mother', 'rather', 'however', 'her', 'better', 'her', 'ever', 'over', 'over', 'mother', 'father', 'mother', 'Her', 'her', 'father', 'mother', 'whatever', 'mother', 'after', 'whatever', 'better', 'whatever', 'greater', 'larger', 'richer', 'ever', 'father', 'father', 'whenever', 'farther', 'consider', 'mother', 'her', 'mother', 'Altogether', 'father', 'mother', 'another', 'her', 'however', 'father', 'mother', 'mother', 'Her', 'ever', 'however', 'father', 'whatever', 'father', 'her', 'her', 'other', 'her', 'her', 'her', 'daughter', 'steadier', 'u'

(b) 包含字母m

```
[w for w in text2 if 'm' in w]
```

```
LCY — python — 80×24

[>>> [w for w in text2 if 'm' in w]
[u'family', u'many', u'manner', u'man', u'many', u'companion', u'home', u'family',
u'whom', u'Gentleman', u'comfortably', u'attachment', u'them', u'merely', u'from',
u'from', u'him', u'comfort', u'former', u'marriage', u'man', u'amply', u'mother',
u'him', u'coming', u'marriage', u'him', u'important', u'might', u'them',
u'from', u'small', u'mother', u'remaining', u'moiety', u'gentleman', u'almost',
u'much', u'disappointment', u'from', u'him', u'terms', u'more', u'himself', u'himself',
u'most', u'him', u'most', u'mother', u'means', u'imperfect', u'many', u'from',
u'meant', u'mark', u'them', u'disappointment', u'temper', u'might', u'many',
u'economically', u'sum', u'from', u'almost', u'immediate', u'improvement',
u'coming', u'twelvemonth', u'remained', u'him', u'recommended', u'command', u'mother',
u'family', u'recommendation', u'time', u'promised', u'make', u'them', u'comfortable',
u'much', u'might', u'them', u'man', u'himself', u'married', u'more', u'amiable',
u'woman', u'might', u'made', u'more', u'might', u'made', u'amiable', u'himself',
u'married', u'himself', u'more', u'minded', u'promise', u'meditated', u'himself',
u'himself', u'income', u'remaining', u'mother', u'warmed', u'made', u'him', u'them',
u'handsome', u'make', u'them', u'completely', u'sum', u'many', u'mother', u'come',
u'from', u'moment', u'much', u'woman', u'common', u'must', u'mind', u'romantic',
u'whomsoever', u'immoveable', u'family', u'them', u'comfort', u'determined',
u'judgment', u'mother', u'them', u'mind', u'must', u'imprudence', u'them',
u'mother', u'many', u'moderation', u'amiable', u'resemblance', u'mother',
u'them', u'themselves', u'admitting', u'mother', u'similar', u'similar',
u'humored', u'imbibed', u'romance', u'much', u'more', u'mistress', u'mother',
u'much', u'himself', u'them', u'some', u'home', u'remaining', u'accommodate',
```

(c) 包含字母序列ph

```
[w for w in text2 if 'ph' in w]
```

```
LCY — python — 80x24

[>>> [w for w in text2 if 'ph' in w]
[u'nephew', u'nephew', u'nephew', u'phrase', u'atmosphere', u'philosophy', u'philanthropic', u'alphabet', u'atmosphere', u'triumph', u'triumph', u'orphan', u'emphasis', u'philippic', u'triumph', u'paragraph', u'triumph', u'triumphantly', u'triumph', u'triumph', u'phrase', u'metaphor', u'Sophia', u'Sophia', u'triumph', u'triumph', u'philosophic', u'paragraph', u'prophecies']]

[>>> [w for w in text2 if w.istitle()]
[u'Sense', u'Sensibility', u'Jane', u'Austen', u'The', u'Dashwood', u'Sussex', u'Their', u'Norland', u'Park', u'The', u'But', u'Mr', u'Henry', u'Dashwood', u'Norland', u'In', u'Gentleman', u'His', u'The', u'Mr', u'Mrs', u'Henry', u'Dashwood', u'By', u'Mr', u'Henry', u'Dashwood', u'The', u'By', u'To', u'Norland', u'Their', u'The', u'He', u'Mr', u'Dashwood', u'The', u'Norland', u'He', u'Mr', u'Dashwood', u'But', u'He', u'His', u'Mr', u'Dashwood', u'Mr', u'John', u'Dashwood', u'His', u'Mr', u'John', u'Dashwood', u'He', u'Had', u'But', u'Mrs', u'John', u'Dashwood', u'When', u'He', u'The', u'Yes', u'It', u'Three', u'He', u'No', u'Mrs', u'John', u'Dashwood', u'No', u'Mrs', u'Dashwood', u'Mrs', u'John', u'Dashwood', u'So', u'Mrs', u'Dashwood', u'Elinor', u'Mrs', u'Dashwood', u'She', u'Marianne', u'Elinor', u'She', u'She', u'The', u'Elinor', u'Mrs', u'Dashwood', u'They', u'The', u'They', u'Elinor', u'She', u'Margaret', u'Marianne', u'Mrs', u'John', u'Dashwood', u'Norland', u'As', u'He', u'Norland', u'Mrs', u'Dashwood', u'A', u'In', u'But', u'Mrs', u'John', u'Dashwood', u'To', u'She', u'How', u'And', u'Miss', u'Dashwoods', u'It', u'Harry', u'It', u'I', u'He', u'I', u'Had', u'He', u'Fanny', u'Perhaps', u'He', u'I', u'But', u'I', u'I', u'The', u'Something', u'Norland', u'Well', u'Consider', u'Your', u'If', u'Why', u'The', u'Harry', u'If', u'To', u'P
```


(d) 除了首字母外是全部小写字母的词 (即titlecase)

```
[w for w in text2 if w.istitle()]
```

```
LCY — python — 80×24

[>>> [w for w in text2 if w.istitle()]]
[u'Sense', u'Sensibility', u'Jane', u'Austen', u'The', u'Dashwood', u'Sussex', u
'Their', u'Norland', u'Park', u'The', u'But', u'Mr', u'Henry', u'Dashwood', u'No
rland', u'In', u'Gentleman', u'His', u'The', u'Mr', u'Mrs', u'Henry', u'Dashwood
', u'By', u'Mr', u'Henry', u'Dashwood', u'The', u'By', u'To', u'Norland', u'Thei
r', u'The', u'He', u'Mr', u'Dashwood', u'The', u'Norland', u'He', u'Mr', u'Dashw
ood', u'But', u'He', u'His', u'Mr', u'Dashwood', u'Mr', u'John', u'Dashwood', u
'His', u'Mr', u'John', u'Dashwood', u'He', u'Had', u'But', u'Mrs', u'John', u'Das
hwood', u'When', u'He', u'The', u'Yes', u'It', u'Three', u'He', u'No', u'Mrs', u
'John', u'Dashwood', u'No', u'Mrs', u'Dashwood', u'Mrs', u'John', u'Dashwood', u
'So', u'Mrs', u'Dashwood', u'Elinor', u'Mrs', u'Dashwood', u'She', u'Marianne',
u'Elinor', u'She', u'She', u'The', u'Elinor', u'Mrs', u'Dashwood', u'They', u'Th
e', u'They', u'Elinor', u'She', u'Margaret', u'Marianne', u'Mrs', u'John', u'Das
hwood', u'Norland', u'As', u'He', u'Norland', u'Mrs', u'Dashwood', u'A', u'In',
u'But', u'Mrs', u'John', u'Dashwood', u'To', u'She', u'How', u'And', u'Miss', u
'Dashwoods', u'It', u'Harry', u'It', u'I', u'He', u'I', u'Had', u'He', u'Fanny',
u'Perhaps', u'He', u'I', u'But', u'I', u'I', u'The', u'Something', u'Norland', u
'Well', u'Consider', u'Your', u'If', u'Why', u'The', u'Harry', u'If', u'To', u'P
erhaps', u'Five', u'Oh', u'What', u'And', u'But', u'I', u'One', u'No', u'I', u'T
here', u'Certainly', u'I', u'I', u'As', u'To', u'They', u'If', u'That', u'I', u'
I', u'My', u'A', u'His', u'To', u'But', u'Mrs', u'Dashwood', u'Fifteen', u'Fanny
', u'Certainly', u'An', u'You', u'I', u'Twice', u'My', u'Her', u'It', u'I', u'I'
, u'It', u'Mr', u'Dashwood', u'One', u'To', u'Undoubtedly', u'They', u'If', u'I'
, u'I', u'I', u'It', u'I', u'I', u'It', u'A', u'I', u'To', u'Indeed', u'I', u'Th
```